



# Analysis of (cDNA) Microarray Data: Part IV. Mixed-Model Equations I



## Mixed-Model Equations

### Setting the scene (1/3):

Kerr & Churchill, 2001  
PNAS 98:8961-8965

Statistical Model:

$$y_{ijkgr} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{igr} + (DG)_{kg} + \epsilon_{ijkgr}$$

*Spots!*  
*Array* *Dye* *Variety* *Gene* *Variety-by-Gene effects* *Gene-specific dye effects*

We assume that there is independent, random error  $\epsilon_{ijkgr}$  with mean 0.

Quantities of interest are expression levels of gene specifically attributable to different varieties:

$$(VG)_{kg} - (VG)_{k'g}$$



### Mixed-Model Equations

Setting the scene (2/3):

Wolfinger et al, 2001  
J Comp Biol 8:625

Two-Step Mixed-Model

1<sup>st</sup> Step: GLOBAL NORMALIZATION

$$\log(y_{ijklm}) = \mu + A_i + D_j + T_k + G_l + TG_{kl} + DG_{jl} + AG_{ilm} + \epsilon_{ijklm}$$

Assumes most genes are not DE. Otherwise some important effects are lost.

$$\log(y_{ijklm}) = \mu + A_i + D_j + T_k + \epsilon_{ijklm}$$

2<sup>nd</sup> Step: GENE MODELS

One for each gene!

Predicted Residuals:  $\hat{\epsilon}_{ijklm} = (y_{ijklm} - \hat{y}_{ijklm})$

$$\hat{\epsilon}_{ijklm} = \mu_l + S_{ilm} + D_{jl} + T_{kl} + \epsilon_{ijklm}$$

Gene-specific treatment effects

$$\epsilon_{ijklm} \sim N(0, \sigma_1^2)$$

Source: G Rosa 2003.

A. Reverter - Sept. 2006, UAB, Barcelona, Spain



### Mixed-Model Equations

Setting the scene (2/3):

Wolfinger et al, 2001  
J Comp Biol 8:625

Two-Step Mixed-Model

Step 1. Global Normalisation

```
proc mixed data=micro covtest;
class array trt dye;
model y=dye trt / out=pred;
random array;
```

Step 1. Gene Models

```
proc mixed data=new covtest;
by clone;
class array trt dye spot;
model resid=dye trt;
random spot(array);
lsmeans trt / diff cl;
ods output diffs=tdiff lsmeans=estimat;
run;
```

Source: G Rosa 2003.

A. Reverter - Sept. 2006, UAB, Barcelona, Spain



## Mixed-Model Equations

### Setting the scene (2/3):

Wolfinger et al, 2001  
J Comp Biol 8:625

#### *Two-Step Mixed-Model*

- 1 Why a two-step approach?  
It is virtually impossible to fit all terms simultaneously in SAS.
- 2 The second approach models gene-specific variance components.
- 3 Residuals from the normalization model are correlated by construction, and yet they are modeled with independent errors in the gene models.  
Little to no difference in practice (Wolfinger et al, 2001).
- 4 Normality in the log-scale.  
Usual assumption; but standard graphical and statistical checks using residuals from gene models should be performed.

Source: G Rosa 2003.

A. Reverter - Sept. 2006, UAB, Barcelona, Spain



## Mixed-Model Equations

### Setting the scene (3/3):

Hoeschele & Li, 2005  
Biostatistics 6:183

#### *Joint versus Gene-Specific Mixed-Models*

1. Flexibility on how the (co)variance structure is modelled and different formulations can be compared, eg. Homo- vs Hetero-geneous within-gene variance)
2. Allows the evaluation of gene-specific treatment contrasts that include main effects
3. Allows model evaluation and residual analysis

1. Low Power (fewer degrees of freedom), but exact if:
2. Genes have the same number of probes in each array;
3. Residual variance must be homogeneous across genes; and
4. Large number of genes.

A. Reverter - Sept. 2006, UAB, Barcelona, Spain



Mixed-Model Equations

Examples: Joint vs Gene-Specific Mixed-Models

Joint

Biol Reprod. 2005 Mar;72(3):546-55. Epub 2004 Oct 13.

Related Articles, Links

Full text article at [www.biolreprod.org](http://www.biolreprod.org)

Identification of differentially expressed genes in individual bovine preimplantation embryos produced by nuclear transfer: improper reprogramming of genes required for development.

Pfister-Genskow M, Myers C, Childs LA, Lacson JC, Patterson T, Betthausen JM, Goneleke PJ, Koppang RW, Lange G, Fisher P, Watt SR, Forsberg EJ, Zheng Y, Leno GH, Schultz RM, Liu B, Chetia C, Yang X, Hoeschele I, Dilertsen KJ.

$$d_i = GV_{i1} - GV_{i2} \text{ vs } d_i = (V_1 + GV_{i1}) - (V_2 + GV_{i2})$$

AI vs IVF vs NT Embryos

Variety has been fitted as an additional fixed effect

Gene-Specific (Two-Step)

Proc Natl Acad Sci U S A. 2005 Dec 6;102(49):17582-7. Epub 2005 Nov 28.

Related Articles, Links

Full Text Article at [www.pnas.org](http://www.pnas.org)

Global gene expression profiles reveal significant nuclear reprogramming by the blastocyst stage after cloning.

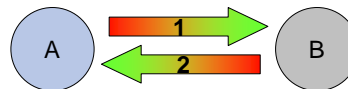
Smith SL, Everts RE, Tian XC, Du F, Sung LY, Rodriguez-Zas SL, Jeong BS, Renard JP, Lewin HA, Yang X.



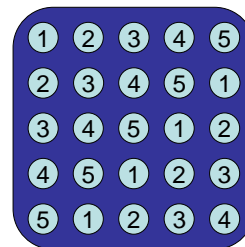
Mixed-Model Equations

Illustration:

2 Arrays, 5 Genes/Array, Genes spotted five times, 2 Treatments (A, B)



array	gene	dye	treat	intens
1	1	R	A	13
1	1	R	B	10
1	2	R	A	11
1	2	R	B	7
1	3	R	A	9
1	3	R	B	9
1	4	R	A	9
1	4	R	B	10
1	5	R	A	4
1	5	R	B	11
2	1	G	A	9
2	1	G	B	9
2	2	G	A	11
2	2	G	B	10
2	3	G	A	11
2	3	G	B	10
2	4	G	A	6
2	4	G	B	10
2	5	G	A	11
2	5	G	B	11



This could represent a row (or column) of the second array.

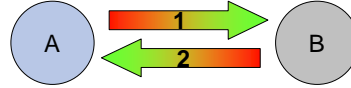
x 5  
(to generate entire data)



### Mixed-Model Equations

#### Illustration:

2 Arrays, 5 Genes/Array,  
Genes spotted five times,  
2 Treatments (A, B)



Model:

$$Y = \text{Array} + \text{Dye} + \underbrace{\text{Gene} \times \text{Treatment}}_{\text{Random}} + \text{Error}$$

array	gene	dye	treat	intens
1	1	R	A	13
1	1	G	B	10
1	2	R	A	11
1	2	G	B	7
1	3	R	A	9
1	3	G	B	9
1	4	R	A	10
1	4	G	B	4
1	5	R	A	11
1	5	G	B	9
2	1	R	A	9
2	1	G	B	11
2	2	R	A	10
2	2	G	B	11
2	3	R	A	10
2	3	G	B	6
2	4	R	A	10
2	4	G	B	11
2	5	R	A	11
2	5	G	B	11

x 5

(to generate entire data)

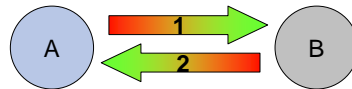
SAS Code:

```
DATA test;
INFILE 'h:\UAB_data\Test\c1c2';
INPUT array gene dye $ treat $ intens;
RUN;
PROC GLM;
CLASS array dye gene treat;
MODEL intens = array dye gene*treat;
RANDOM gene*treat;
LSMEANS array dye gene*treat / pdiff;
RUN;
```



### Mixed-Model Equations

#### Illustration:



Dependent Variable: intens

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	130.7500000	11.8863636	4.12	<.0001
Error	88	254.0000000	2.8863636		
Corrected Total	99	384.7500000			

R-Square	Coeff Var	Root MSE	intens Mean
0.339831	17.78984	1.698930	9.550000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
array	1	6.2500000	6.2500000	2.17	0.1447
dye	1	2.2500000	2.2500000	0.78	0.3797
gene*treat	9	122.2500000	13.5833333	4.71	<.0001

$$\sigma_e^2 = 1.70^2 = 2.89$$

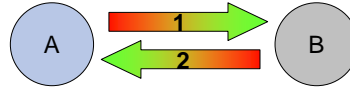
$$13.58 = \sigma_e^2 + 10\sigma_{G \times T}^2 \Rightarrow \sigma_{G \times T}^2 = 1.07$$

$$\frac{1.07}{1.07 + 2.89} = 0.27 \Rightarrow 1.35 \text{ Genes!}$$



Mixed-Model Equations

Illustration:



gene	treat	intens LSMEAN	LSMEAN Number
1	A	11.0000000	1
1	B	9.5000000	2
2	A	10.5000000	3
2	B	9.0000000	4
3	A	9.5000000	5
3	B	10.0000000	6
4	A	9.5000000	7
4	B	8.0000000	8
5	A	7.5000000	9
5	B	11.0000000	10

Least Squares Means for effect gene\*treat  
Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: intens

i/j	1	2	3	4	5	6	7	8	9	10
1		0.0515	0.5122	0.0100	0.0515	0.1915	0.0515	0.0002	<.0001	1.0000
2	0.0515		0.1915	0.5122	1.0000	0.5122	1.0000	0.0515	0.0100	0.0515
3	0.5122	0.1915		0.0515	0.1915	0.5122	0.1915	0.0014	0.0002	0.5122
4	0.0100	0.5122	0.0515		0.5122	0.1915	0.5122	0.1915	0.0515	0.0100
5	0.0515	1.0000	0.1915	0.5122		0.5122	1.0000	0.0515	0.0100	0.0515
6	0.1915	0.5122	0.5122	0.1915	0.5122		0.5122	0.0100	0.0014	0.1915
7	0.0515	1.0000	0.1915	0.5122	1.0000	0.5122		0.0515	0.0100	0.0515
8	0.0002	0.0515	0.0014	0.1915	0.0515	0.0100	0.0515		0.5122	0.0002
9	<.0001	0.0100	0.0002	0.0515	0.0100	0.0014	0.0100	0.5122		<.0001
10	1.0000	0.0515	0.5122	0.0100	0.0515	0.1915	0.0515	0.0002	<.0001	

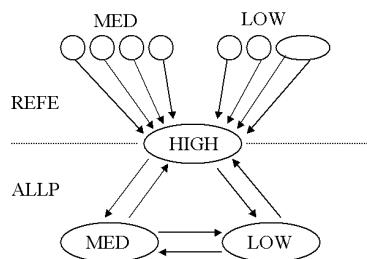
A. Reverter - Sept. 2006, UAB, Barcelona, Spain



Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005  
J Anim Sci 83:1-12



Log<sub>2</sub> Intensities

$$y_{ijkgr} = C_i + A_j + D_k + T_t + (AD)_{jk} + G_g + (AG)_{jg} + (DG)_{kg} + (TG)_{tg} + \epsilon_{ijkgr}$$

- Component of Design (Reference, All-Pairs)
  - Array slide (1, 2, ..., 14)
  - Dye (Red, Green)
  - Treatment (Diets: High, Medium, Low)
  - Array \* Dye
  - Main effect of Gene
  - Gene \* Array
  - Gene \* Dye
  - Gene \* Treatment
  - Random Error
- } Random effects

A. Reverter - Sept. 2006, UAB, Barcelona, Spain



Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005  
J Anim Sci 83:1-12

1. The terms  $C$ ,  $A$ ,  $D$ ,  $T$ , and  $AD$  account for all effects that are not gene specific, have no biological significance, and the fitting of which aims at normalizing the data by accounting for systematic effects.
2. The random gene effect  $G$  contains the average level of gene expression (averaged over the other factors).
3. The random gene  $\times$  array in  $(AG)$  models the effects for each spot and it serves to account for the spot-to-spot variability inherent in spotted microarray data. It allows us to extract appropriate information about the treatments and obviates the need to form ratios (Wolfinger et al., 2001).
4. The random gene  $\times$  dye in  $(DG)$  models the gene specific dye effects occurring when some genes exhibit higher fluorescent signal when labeled with one dye or the other, regardless of the treatment (Kerr et al., 2002).
5. The effect of interest was the random interaction between genes and diet treatments,  $(TG)$  because it captured differences from overall averages that were attributable to specific combination of diet and gene.

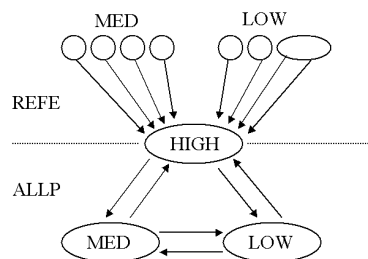
A. Reverter - Sept. 2006, UAB, Barcelona, Spain



Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005  
J Anim Sci 83:1-12



Log<sub>2</sub> Intensities

$$y_{ijkgr} = C_i + A_j + D_k + T_t + (AD)_{jk} + G_g + (AG)_{jg} + (DG)_{kg} + (TG)_{tg} + \epsilon_{ijkgr}$$

Main effect of Gene  
Gene \* Array  
Gene \* Dye  
Gene \* Treatment  
Random Error

Decomposition of Total Variance

- Between Genes
- B/w G within Array
- B/w G within Dye
- B/w G within Treat.
- Within Gene

A. Reverter - Sept. 2006, UAB, Barcelona, Spain



Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005  
J Anim Sci 83:1-12

Log<sub>2</sub> Intensities

$$y_{ijkgr} = C_i + A_j + D_k + T_t + (AD)_{jk} + G_g + (AG)_{jg} + (DG)_{kg} + (TG)_{tg} + \epsilon_{ijkgr}$$

```

areverte@bioserver: ~/BeefCRC/Keren.AG2JAS
General statistics:
traits # rec. min. max. avg. std.
expr 300936 .00000 15.99830 9.50480 2.17957
    
```

```

areverte@bioserver: ~/BeefCRC/Keren.AG2JAS
*****
* MODEL INFORMATION *
*****
Factor T nested # skp expr
-----
comp F 2 x
array F 14 x
dye F 2 x
trt F 3 x
axd F 28 x
gene R 7347 x
gxa R 75443 x
gxd R 14694 x
gxt R 22041 x
    
```

A. Reverter - Sept. 2006, UAB, Barcelona, Spain



Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005  
J Anim Sci 83:1-12

REML Estimates

Model	Variance component <sup>a</sup>					LogL
	$\sigma^2_{\epsilon}$	$\sigma^2_g$	$\sigma^2_{jg}$ Gene * Array	$\sigma^2_{kg}$ Gene * Dye	$\sigma^2_{tg}$ Gene * Diet	
Full	0.238	3.656	0.536	0.023	0.164	182,732
R1	0.719	3.662	-	0.002	0.137	105,449
R2	0.244	3.686	0.535	-	0.170	181,933
R3	0.720	3.664	-	-	0.137	105,444

%Total Variance

Full	5.1	79.2	11.6	0.5	3.6
------	-----	------	------	-----	-----

A. Reverter - Sept. 2006, UAB, Barcelona, Spain

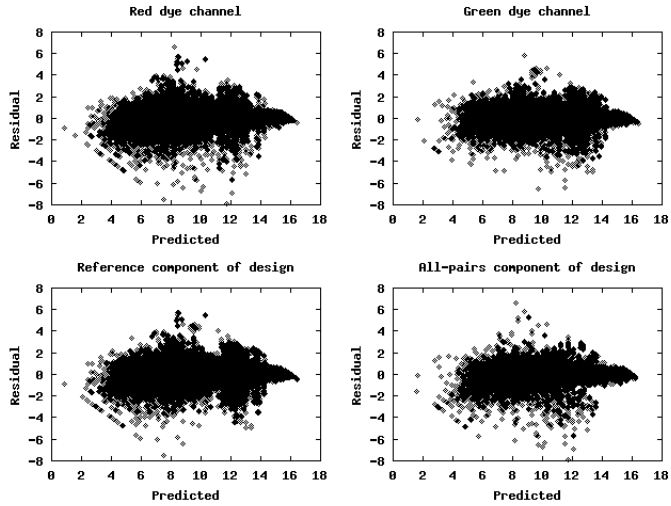


### Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005  
J Anim Sci 83:1-12

Residual Plots



A. Reverter – Sept. 2006, UAB, Barcelona, Spain



### Mixed-Model Equations

The Diets experiment:

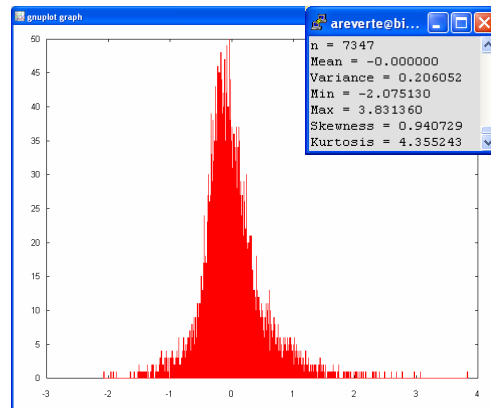
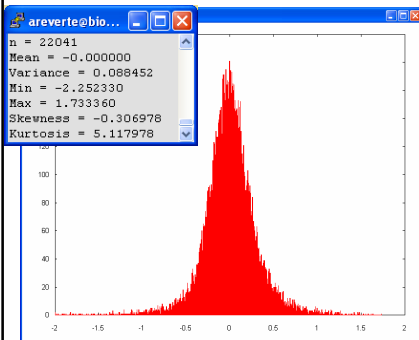
Byrne et al, 2005. J Anim Sci 83:1-12

Measures of (possible) Differential Expression

$(\hat{T}G)$  BLUP Solutions

HIGH vs LOW Contrast

$$d_g = (\hat{T}G)_{HIGH,g} - (\hat{T}G)_{LOW,g}$$



Alternatively:

$$d_g^* = (\hat{T}G)_{HIGH,g} - (\hat{T}G)_{MED,g}$$

$$d_g^* = (\hat{T}G)_{MED,g} - (\hat{T}G)_{LOW,g}$$

A. Reverter – Sept. 2006, UAB, Barcelona, Spain



Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005. J Anim Sci 83:1-12

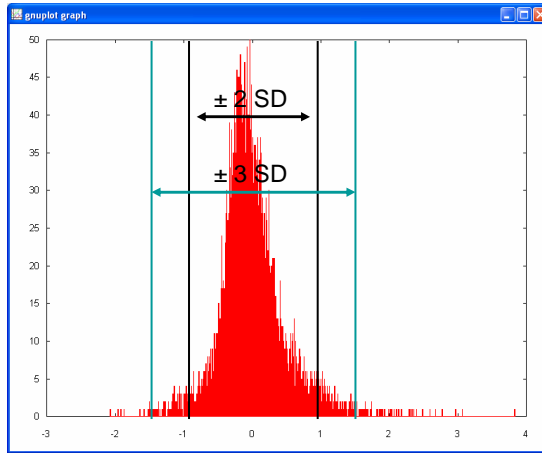
%Total Variance

Full	5.1	79.2	11.6	0.5	3.6
------	-----	------	------	-----	-----

$$d_g = (\hat{TG})_{HIGH,g} - (\hat{TG})_{LOW,g}$$

(SD = 0.454)

$d_g$	$\pm 2$ SD	$\pm 3$ SD
< 0	129	29
> 0	309	82
<b>% Total</b>	<b>5.8</b>	<b>1.5</b>



A. Reverter - Sept. 2006, UAB, Barcelona, Spain

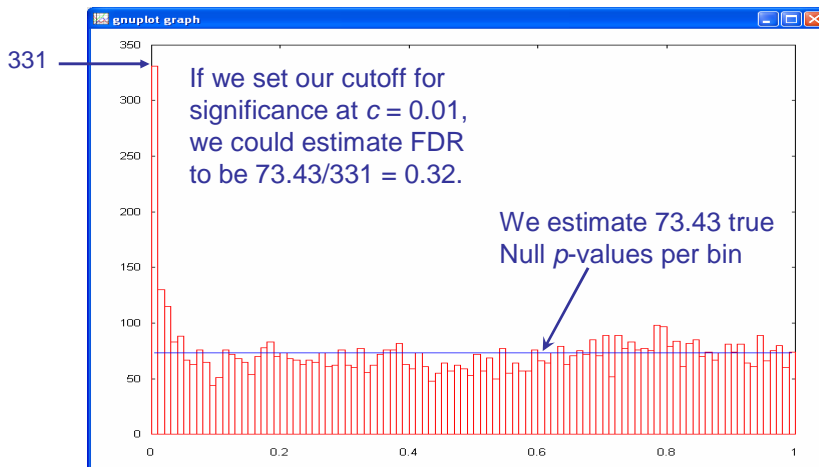


Mixed-Model Equations

The Diets experiment:

Byrne et al, 2005. J Anim Sci 83:1-12

Histogram of  $p$ -values for a Test of  $d_g$  assuming Normality  
(Using 100 bins at 0.01 interval)



A. Reverter - Sept. 2006, UAB, Barcelona, Spain



## Mixed-Model Equations

### Concluding Remarks on DE Genes:

1. The assumption of Normality of  $d_g$  is questionable
2. Hence, resorting to “tabled values” (ie. 2, 3, SD from the mean) could be suboptimal
3. Instead, using the proportion of the total variation that is attributed to the Gene by Treatment (diet) interaction could be safer option.
4. We let the mixed-model tell us how many genes are likely to be DE
5. CLAIM: The proportion of the total variation that is attributed to the Gene by Treatment (diet) interaction allows us to control the FDR
6. CLAIM: The proportion of the total variation that is attributed to the Gene by Treatment (diet) interaction provides a lower bound for the mixing proportion in the extreme cluster(s) in a model based clustering via mixtures of distributions.